

BIONUMERICS Tutorial: wgSNP: analysis and clustering

1 Introduction

A Single Nucleotide Polymorphism (SNP) is a variation in a single nucleotide, which occurs at a specific position of the genome. When performed on whole genome sequences (WGS), this analysis is referred as **whole genome SNP (wgSNP) analysis**. When performed in BIONUMERICS a typical workflow for wgSNP analysis consists of following steps:

- 1. Choose a reference sequence
- 2. Map sequence reads against the reference sequence (locally or on the cloud calculation engine)
- 3. Perform wgSNP analysis and filter out relevant SNPs
- 4. wgSNP clustering

In this tutorial we will focus on the last two steps of the workflow. The first two steps are covered in the tutorials "wgSNP with mapping performed on the cloud calculation engine" and "wgSNP with mapping performed locally on your own computer".

2 Preparing the database

The **WGS demo database** for *Staphylococcus aureus* contains entries for which a reference mapping has already been calculated. The resulting sequences are stored in the **SNP outbreak** sequence type and will be used in this tutorial to illustrate the *SNP filtering* window, the use of SNP filters, and possible follow-up analysis tools available in BIONUMERICS.

This demo database can be downloaded directly from the *BIONUMERICS Startup* window (see 2.1), or restored from the back-up file available on our website (see 2.2).

2.1 Option 1: Download demo database from the Startup Screen

1. To download the database directly from the *BIONUMERICS Startup* window, click the button, located in the toolbar in the *BIONUMERICS Startup* window.

This calls the *Tutorial databases* window (see Figure 1).

🖆 Tutorial databases					- 0
ile Database Tutorial Window					
Futorial databases					
Name	Size (MB)	Min	Max	-	
BandScoring demo database	15	7	8	^	The DemoBase Connected demonstration
DemoBase Connected	4	7.5	8		database in BioNumerics version 8.0 format.
Neisseria MLST demo database	1	7.5	8		This database is used in several tutorials and comes with the BioNumerics installation.
SNP demonstration database	0	7.5	8		comes with the biowdiffences installation.
WGS demo database for Listeria monocytogenes	311	7.5	8		
WGS_demo_database_for_Brucella_spp	214	7.6	8		
WGS_demo_database_for_Escherichia_coli	601	7.5	8		
WGS_demo_database_for_Staphylococcus_aureus	624	7.5	8		
				~	
	<			>	
					-
Tutorials					
D .					
Name				-	
Band matching and polymorphism analysis				^	This tutorial illustrates how to calculate a
Building automated decision and action workflows					Principal Components Analysis (PCA) and a Multi Dimensional Scaling (MDS)
Calculating a PCA and an MDS on a fingerprint data set					(sometimes also called Principal
					Coordinates Analysis (PCoA)) on a
Clustering a phenotypic test assay					fingerprint data set and how to change the
Combined analysis of character data					layout of the obtained plots.
Combined analysis of character data Combined analysis of fingerprint data					layout of the obtained plots.
Clustering a phenotypic test assay Combined analysis of character data Combined analysis of fingerprint data Configuring the database layout Dendrogram layout options					layout of the obtained plots.

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

- 2. Select the WGS_demo_database_for_Staphylococcus_aureus from the list and select Database > Download ().
- 3. Confirm the installation of the database and press < OK > after successful installation of the database.
- 4. Close the *Tutorial databases* window with *File* > *Exit*.

The **WGS_demo_database_for_Staphylococcus_aureus** appears in the *BIONUMERICS Startup* window.

5. Double-click the **WGS_demo_database_for_Staphylococcus_aureus** in the *BIONUMERICS Startup* window to open the database.

2.2 Option 2: Restore demo database from back-up file

A BIONUMERICS back-up file of the whole genome demo database for Staphylococcus aureus is also available on our website. This backup can be restored to a functional database in BIONU-MERICS.

6. Download the file wgMLST_SAUR.bnbk file from https://www.applied-maths.com/download/ sample-data, under 'WGS_demo_database_for_Staphylococcus_aureus'.



In contrast to other browsers, some versions of Internet Explorer rename the wgMLST_SAUR.bnbk database backup file into wgMLST_SAUR.zip. If this happens, you should manually remove the .zip file extension and replace with .bnbk. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the .zip file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

- 7. In the *BIONUMERICS Startup* window, press the button. From the menu that appears, select **Restore database...**.
- 8. Browse for the downloaded file and select *Create copy*. Note that, if *Overwrite* is selected, an existing database will be overwritten.
- 9. Specify a new name for this demonstration database, e.g. "Whole genome Staphylococcus aureus demobase".
- 10. Click < OK > to start restoring the database from the backup file (see Figure 2).

C:\	Users\Downloads\v	gMLST_SAUR1.b	nbk		Browse
Dest	ination				
	Overwrite				
0	Overwrite the curr coli with a previous since the backup v recovered.	s version of the da	ta as stored in a	a backup. An	y new data
۲	Create a new data have access to bo copy a database fr	th old and current	versions of an o		
Mau	v database name:	Whole genome	Stanhvinco		

Figure 2: Restoring the whole genome demonstration database from the BioNumerics backup file wgMLST_SAUR.bnbk.

11. Once the process is complete, click < *Yes*> to open the database.

The *Main* window is displayed (see Figure 3).

3 Perform wgSNP analysis and filter out revelant SNPs

In our demonstration database, a reference mapping has already been calculated for the entries from the **Neonatal MRSA study**. The resulting sequences are stored in the **SNP outbreak** sequence type and will be used in this section to start a SNP analysis and to illustrate the use of SNP filters in the *SNP filtering* window.

- 1. Make sure no selection is present in the *Database entries* panel: press the **F4**-key to clear any entry selection.
- 2. In the *Database entries* panel, select the **Neonatal MRSA study** view from the list (see Figure 4) and use *Edit* > *Select all* (Ctrl+A) to select all 14 entries contained in this study.

WGS demo database for Staphy	lococcus aureus - Bi	oNumerics								- 0
Edit Database Analysis			elp							
	h. D Q	\$0 0,								
		φ •		_						
periment types				Dat	abase entries					Comparisons
A + P ⊗ €		: †	L <all< th=""><th>E</th><th>1 + 🖻</th><th>⊗ 🛃 🗗</th><th>All Entries></th><th>10</th><th></th><th>+ 13 8 8 1 6 2 4</th></all<>	E	1 + 🖻	⊗ 🛃 🗗	All Entries>	10		+ 13 8 8 1 6 2 4
# Name	Туре				Kou	Organism name	Study title	ST info	outbreak 🕶 1 2 3 4 5 6 7 8	Name Modified date
					NOY					
≥ 1 wgs 2 wgMLST	Sequence read Character type		í l	1 8	ERR103403	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin A neonatal MRSA outbreak investigation usin		part of outbrea	Neonatal MRSA study 2015-01-28 11:14:43 Plot study 2015-01-28 12:15:55
					ERR103402 ERR103398	Staphylococcus aureus				WGS for MRSA outbreak 2015-01-28 12:15:55
3 denovo	Sequence type				ERR103398 ERR103394	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin A neonatal MRSA outbreak investigation usin		during outbreat • • • • • • • • • • • • • • • • • • •	WUS for MRSA outpreak 2015-01-28 12:23:44
4 quality ≩ 5 wgs TrimmedStats	Character type				ERR103394 ERR101899	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin. A neonatal MRSA outbreak investigation usin.		part of outbrea	
6 wgMLST_CalTypes					ERR101899	Staphylococcus aureus Staphylococcus aureus	A neonatal MRSA outbreak investigation usin. A neonatal MRSA outbreak investigation usin.		prior to outbrea	
7 SNP outbreak	Sequence type			ΗH	ERR103395	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.		prior • • • • • •	
F 7 Sive Outbreak ≩ 8 wosLong	Sequence read				ERR103395	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.		part of outbreal	<
-e o wyscony	Sequence read	set types			ERR103405	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.		part of outbrea	
				' ö	ERR101900	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.		part of outbreal • • • • • • •	Identification projects Decision networks
Fields Database design				Ö	ERR103401	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.		part of outbrea	29 + 13 ⊗ 18, 18, 5.
					ERR159680	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.		during outbreak • • • • • •	
⊢ฅ⊗฿∣ฅ	∽↑	↓ <all< td=""><td>Entry fields></td><td></td><td>ERR103397</td><td>Staphylococcus aureus</td><td>A neonatal MRSA outbreak investigation usin.</td><td>1</td><td>during outbreak • • • • • •</td><td>Name Modified date</td></all<>	Entry fields>		ERR103397	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.	1	during outbreak • • • • • •	Name Modified date
Name	Field type	1		ī	ERR103404	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin.	22	part of outbrea	
	Fixed				ERR127429	Staphylococcus aureus			S. aureus clust	
	Fixed		,		ERR127426	Staphylococcus aureus	A pilot study of rapid benchtop sequencing		S. aureus clust • • • • • •	
	Fixed			Π	ERR127428	Staphylococcus aureus			S. aureus clust • • • • • •	
	Fixed				ERR127417	Staphylococcus aureus	A pilot study of rapid benchtop sequencing		S. aureus clust	
	Fixed				ERR127408	Staphylococcus aureus			S. aureus clust	
	Fixed				ERR127443	Staphylococcus aureus	A plot study of rapid benchtop sequencing		S. aureus clust	
	Fixed				ERR127448	Staphylococcus aureus			S. aureus clust	
	Fixed				ERR127432	Staphylococcus aureus			S. aureus clust	Alignments BLAST projects Chromosome comparison
	Flexible				ERR127431	Staphylococcus aureus	A plot study of rapid benchtop sequencing		S, aureus clust	
	Flexible				ERR127412	Staphylococcus aureus	A plot study of rapid benchtop sequencing		S. aureus clust • • • • • •	+ 🖻 🖉 🗟 🗉 🔍 斗
mean rundlar at	1 NAMES				ERR127427	Staphylococcus aureus			S. aureus clust • • • • • •	Name Modified date
					ERR127444	Staphylococcus aureus	A pilot study of rapid benchtop sequencing		S. aureus clust • • • • • •	
					ERR127413	Staphylococcus aureus			S. aureus clust • • • • • •	
rprint files Power assemblies	Annotations				ERR127442	Staphylococcus aureus			S. aureus clust • • • • • •	
} + P→ ⊗ B₀	fi v		and Press	1 0	ERR127441	Staphylococcus aureus			S. aureus clust • • • • • •	
	- CD	<all finger<="" td=""><td>orint ties></td><td></td><td>ERR127409</td><td>Staphylococcus aureus</td><td></td><td></td><td>S. aureus clust • • • • • •</td><td></td></all>	orint ties>		ERR127409	Staphylococcus aureus			S. aureus clust • • • • • •	
File name E					ERR127433	Staphylococcus aureus	A pilot study of rapid benchtop sequencing		S. aureus clust • • • • • •	
			3		ERR127415	Staphylococcus aureus	A plot study of rapid benchtop sequencing		S. aureus clust • • • • • •	
<					<				> < >	

Figure 3: The Staphylococcus aureus demonstration database: the Main window.

Database entries	3			
£ + C	9 ⊗ & ₽	▽	<all entries=""> し</all>	
Key	Organism name	Study title	<loaded entries=""></loaded>	4 5
ERR070033	Staphylococcus aureus	WGS for an	All Entries>	• •
ERR070034	Staphylococcus aureus	WGS for an		
ERR070035	Staphylococcus aureus	WGS for an	<my entries=""></my>	
ERR070036	Staphylococcus aureus	WGS for an	<selected entries=""></selected>	[•••]
ERR070037	Staphylococcus aureus	WGS for an		
ERR070038	Staphylococcus aureus	WGS for an	< Manage user defined views>	 • • •
ERR070039	Staphylococcus aureus	WGS for an		
ERR070040	Staphylococcus aureus	WGS for an	Neonatal MRSA study	••
ERR070041	Staphylococcus aureus	WGS for an	Pilot study	
ERR070042	Staphylococcus aureus	WGS for an	arysis of meticillin-res Saur outo 23/1	4 a - a - a

Figure 4: Select a view from the list.

- 3. Select *Analysis* > *Sequence types* > *Start SNP analysis...* to start the *SNP analysis* wizard.
- 4. Select **SNP outbreak** as *Experiment type* and press <*Next*> (see Figure 5).

A number of predefined SNP templates are available.

- 5. Highlight the *Strict filtering* template and press <*Next*> (see Figure 6).
- 6. Check Open SNP analysis window and press < Finish> (see Figure 7).

It will take a few moments to load the sequences and apply the filters from the SNP template. The resulting *SNP filtering* window is shown in Figure 8.

This window consists of following panels:

- The *Entries* panel shows all entries that are included in the SNP analysis, with all entry information fields. Two additional fields are present: 'Total' shows the raw number of SNPs (i.e. without any SNP filter applied) and 'Retained' shows the number of SNPs after applying all active SNP filters for the sample sequence.
- The *Filters* panel shows the list of SNP filters that are applied, with the 'Info' column showing additional information regarding the filter and applied settings (if applicable). This list is

			_	
Perform SNP analysis			?	×
Experiment type Choose input experiment type.				
Choose the experiment type of the input data.				
SNP outbreak	•			
NOTE: only reference mapped Sequence experiment types containing data for all selected entries are listed.				
	< Back	Next >	Can	cel

Figure 5: Select the input sequences.

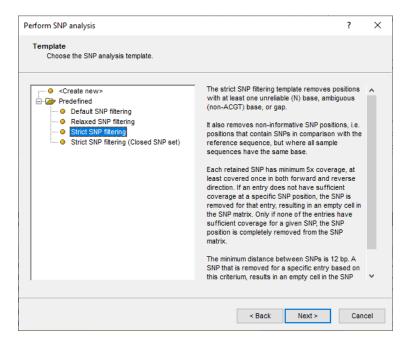


Figure 6: Choose a SNP filtering.

initially populated from the SNP template, but SNP filters can be added or removed and their settings can be changed.

- The *SNP Positions* panel shows information on all positions where at least one SNP was detected. For each SNP filter that is listed in the *Filters* panel, a column is displayed with the filter's result on each position. The bottom of this panel shows a sub-panel with the details on the highlighted position, i.e. showing the base and filter results for all the sample sequences on that position.
- The Entry SNPs panel lists the SNPs for the highlighted entry in the Entries panel.
- The Genome panel shows the SNPs on a genome view.

Analysis Choose to visualize the analysis or directly export to a comparison. • Open SNP analysis window Open the selected template in the SNP analysis window. • New comparison Run the SNP analysis and export the results to a new comparison. • New comparison Run the SNP analysis and export the results to a new comparison. • Sack • Finish Cancel	erfor	m SNP analysis	?	>
Open the selected template in the SNP analysis window. New comparison Run the SNP analysis and export the results to a new comparison.	Ana			
Open the selected template in the SNP analysis window. New comparison Run the SNP analysis and export the results to a new comparison.	۲	Open SNP analysis window		
Run the SNP analysis and export the results to a new comparison.	-			
	0	New comparison		
	Ŭ	Run the SNP analysis and export the results to a new comparison.		
Parameter and a second s				
2 minute and a second se				
		< Back Finish		

Figure 7: Open SNP analysis window.

	oreak (SNP A	nalysis) cks SNPs Windo											-	
ntries	Graph Tra	CKS SINPS WINDO	им нер				Filters							
30							+ ⊗	ļtļ 🗞						
Retained	Total	Key	Organism name	Study title	ST info	•	Filter		Info					
45	155487	ERR103396	Staphylococcus aureus	A neonatal MRSA out	22	^	Inter-SNP	distance	Minimun	12 bp t	etween SNPs.			
1857	1199574	ERR103397	Staphylococcus aureus	A neonatal MRSA out	1		Non-infor	mative SNPs	Remove	non-inf	ormative SNP positions.			
2195	1059597	ERR103398	Staphylococcus aureus	A neonatal MRSA out	1		Absolute	coverage	Total: 5.	Forwar	d: 1. Reverse: 1			
44	123426	ERR103400	Staphylococcus aureus	A neonatal MRSA out	22		Ambiguo	is bases	Remove	position	is with at least one ambig	uous		
0	0	ERR103401	Staphylococcus aureus	A neonatal MRSA out	22		Unreliable				s with at least one unreli			
38	158148	ERR103402	Staphylococcus aureus	A neonatal MRSA out	22		Gaps		Remove	position	is with at least one gap.			
24	73297	ERR103403	Staphylococcus aureus	A neonatal MRSA out	22									
4	73699	ERR103404	Staphylococcus aureus	A neonatal MRSA out	22									
4	71100	ERR103405	Staphylococcus aureus	A neonatal MRSA out	22		SNP Positio	ns						
3614	1258752	ERR159680	Staphylococcus aureus	A neonatal MRSA out	36	~	# F	osition	NPs		Inter-SNP distance	Non-informative	Absolute covera	ge Ambigu
<						>	40	51330	3 → T		0% (0/14) reject	ted 🗸 93% G (13/14) 🗸 0% (0/14) reje	cted 🗸 0%
tries Tracl	ks						41	51480	⇒ C			ted 🖌 71% T (10/14		
							42	51536	A → G		0% (0/14) reject	ted 🗸 93% A (13/14) 🗸 0% (0/14) reje	cted 🖌 0% (
							43		A → G			ted 🗸 86% A (12/14		
	目 ↑						44		A → G			ted 🗸 86% A (12/14		
0							45	53854				ted 🖌 93% G (13/14		
	50	500 54000 545	00 55000 55500 56	000 56500 57000	57500 58000	\sim	46	53863	3 → A		0% (0/14) reject	ted 🗸 86% G (12/14) 🗸 7% (1/14) reje	cted 🖌 0% (
[•] 52500	53000 53	500 54000 545			57500 58000		47	53919			0% (0/14) reject	ted 🗸 93% G (13/14) 🗸 0% (0/14) reje	cted 🗸 0% (
SAUR 28			Ĝau				48	55390				ted 🖌 86% G (12/14		
SAUR 200	13		PA0	PAOR			49	56625) → G		0% (0/14) reject	ted 🖌 79% C (11/14) 🗸 7% (1/14) reje	ected 🖌 0% (
							50	57329	A → G		✓ 0% (0/14) reject	ted 🗸 93% A (13/14) 🗸 0% (0/14) reje	cted 🗸 0% (
		1		1	1.1		51	57431	⇒ T		0% (0/14) reject	ted 🖌 79% C (11/14) 🗸 0% (0/14) reie	ected 🖌 0% (
					1			۲						>
				X			Key		SNE		Inter-SNP distance	Non-informative	Absolute coverage	Ambiguou
					$\langle \rangle$				in the second		Inter-swe distance	Non-informative	Absolute coverage	Ambiguou
				,		\	ERR101899			⇒ G				
				-0.1		1	ERR101900			→ G				
				[50		1	ERR103394		-	⇒ G				
				100		/	ERR103395		G		✓ 92 bp > 12 bp	✓ 86% G (12/14)	✓ 14-8-6	0% (0/1
)		/	ERR103396			⇒ G				
					$\overline{}$		ERR103397			⇒ G				
<					\sim	<u> </u>	ERR103398			⇒ G				>
_	_					,	ERR103400	_	<					>
nome SN	^o matrix						SNP Positions	Entry SN						

Figure 8: SNP analysis window.

- The *Tracks* panel in default view is displayed as a tab with the *Entries* panel. With this panel, you can determine which tracks are plotted in the *Genome* panel.
- The SNP matrix panel shows the resulting SNP matrix, as it would be exported.

Whenever possible, the cursor position is synchronized between the different panels:

7. Click on a position in the *SNP Positions* panel for example.

The details in the bottom part of the panel are updated and so is the *Genome* panel: the graph will show the position. Furthermore, the clicked position in the *SNP Positions* panel will appear

highlighted in the *Entry SNPs* panel, *only* if the currently highlighted entry in the *Entries* panel has a SNP at that position.

8. Double-click a position in the details panel (bottom part of the *SNP Positions* panel) or in the *Entry SNPs* panel.

This action will open the *Sequence editor* window of the corresponding sequence, with this position highlighted. If a sequence assembly is available in BIONUMERICS, the \bowtie will be active and selecting *File* > *Open assembler* (\bowtie) will open the assembly.



Sequence assemblies are not available when the remapping was performed on the calculation engine.

- 9. A SNP filter can be added with *Filters* > *Add filter...* (+).
- 10. Check or uncheck an individual SNP filter in the *Filters* panel to view its effect.

When the toggle *Filters* > *Toggle rejected SNP visibility* is unchecked (\circledast), the positions in the *SNP Positions* panel and the *Entry SNPs* panel will be limited to the retained SNPs, i.e. those SNPs that have passed the applied SNP filters.

When the toggle is checked (()) the listed positions in both panels correspond to the total (i.e., unfiltered) SNP set.

SNP out																														-		×
File Filters Entries	Graph T	racks S	NPs \	Windov	w Hel	p									_	_			_					_	_				_	_		
																	Filte															
盟																	+	- ⊗	ļţĻ <	8												
Retained	Total	Key			Org	ganism	n name		Study				ST info			•		Filter			Info											•
45	155487	ERR	03396		Sta	phyloco	occus a	aureus	A neo	natal M	RSA ou	ıt	22			^		Inter-SNP	distance		Minimur	n 12 bp	betw	een SNPs								,
1857	1199574	ERR	03397		Sta	phyloco	occus a	aureus	A neo	natal M	RSA ou	ıt	1				~	Non-infor	mative SI	NPs	Remove	e non-in	forma	tive SNP (ositions.							
2195	1059597		03398		Sta	phyloco	occus a	aureus	A neo	natal M	RSA ou	ıt	1					Absolute	coverage	e	Total: 5	, Forwa	ard: 1,	Reverse:	1							
44	123426		03400			phyloco					RSA ou		22					Ambiguo						ith at leas								
•	0		03401			phyloco					RSA ou		22					Unreliable	bases					ith at least		liable b.						
38	158148		03402			phyloco					RSA ou		22				≤	Gaps			Remove	e positio	ons w	ith at least	one gap							
24	73297		03403			phyloco					RSA ou		22																			
✓ 4 ✓ 4	73699 71100		03404			phyloco					RSA ou RSA ou		22 22				SNE	P Positio	ns													
3614	1258752		03405 59680			phyloca phyloca			A neo				22 36			J			osition	SNE	20			Inter-SN	D diotop		on-informative		Absolute c		Amb	
<	1230132	LINK	33000		Ju	phyloca	000030	urcus	A 1100	natarini	1.54.00	····				>					-							_				
Entries Trac	due .															-		40	5133		\rightarrow T \rightarrow C						 93% G (13/14 71% T (10/14 		 0% (0/1 0% (0/1 			•
Entries Trac	:KS																	41			⇒ C ⇒ G						93% A (13/14)		 0% (0/1 0% (0/1 			
SNP matrix																		43	5156		⇒ G						86% A (12/14		 0% (0/1 0% (0/1 			
																		44	5224		⇒ G						/ 86% A (12/14		 0% (0/1 			
	75																	45	5385		⇒ A						/ 93% G (13/14		✓ 0% (0/1			
	726		1 2													1		46	5386	G	⇒ A			✓ 0%	(0/14) rej	ected 💊	/ 86% G (12/14)	7% (1/1	14) rejecte	d 🗸	0% (
	5	: :	5	7	7	3	5	5	3	7	7	5	7	1	7			47	5391	9 G	⇒ A			0%	(0/14) rej	ected s	/ 93% G (13/14)	 0% (0/1 	14) rejecte	d 🗸	0% (
Reference	т		С	G	A	т	С	A	A	A	Т	A	С	Т	A	^		48	5539) G	⇒ A			0%	(0/14) rej	ected 💊	/ 86% G (12/14)	 0% (0/1 	14) rejecte	d 🗸	0% (
ERR101899	Т		C C	G	A	Ţ	C C	A	A	A	1	Â	C C	T	A			49	5662	5 C	⇒ G						/ 79% C (11/14		7% (1/1			
ERR101900 ERR103394			c	G	Ā	÷	c	A	Â	Â	÷	<u>^</u>	c	÷	Â			50			⇒ G						93% A (13/14		0% (0/1			
ERR103394 ERR103395			c	G	G	÷	c	c	Å	Â	÷	2	c	c	Â			51		1 C	→ T			✓ 0%	(0/14) rej	ected N	🖊 79% C (11/14)	0% (0/1	 rejecte 	d 🗸	0% (
ERR103396			c	G	A	÷	c	A	Â	Â	÷	A	c	т	Â			52 '														`
ERR103397	T T		c	G	G	Ť	т	Â	A	Â	_	A	т	т	A		Key				SN	Р	Int	er-SNP d	istance	Non	-informative	Ab	solute cove	erage /	Ambigu	iou 🔊
ERR103398	т	A	c	G	G	T	Ť	A	A	A	С	A	T	т	A			101899			G	⇒ G										
ERR103400	т	G	С	G	A	т	С	А	A	A	т	А	С	т	A			101900				⇒ G										
ERR103401	т	G	с	G	А	т	с	А	А	А	т	А	с	т	А			103394			_	⇒ G										
ERR103402	т		с	G	А	т	С	А	А	А	т	А	с	т	А		ERR	103395			G	⇒ A	~	92 bp >	12 bp	~	86% G (12/14)	~	14-8-6		 0% 	(0/1-
ERR103403	т		С	G	A	т	С	A	A	A	т	Α	С	т	A		ERR	103396			G	⇒ G										
ERR103404	т		С	G	A	т	С	A	A	А	т	А	С	т	A		ERR	103397			G	⇒ G										
ERR103405	Т	G	С	G	A	T	С	A	A	A	Т	A	С	т	A	~		103398				⇒ G										
ERR159680	<															>		103400		_	<											>
Genome	P matrix																SNP															

11. Click on the tab of the SNP matrix panel to show the SNP matrix (see Figure 9).

Figure 9: SNP matrix displayed.

12. Select *File* > *Export to comparison...* (🕎) to export the SNP matrix to a comparison.

In the *Comparison* window a cluster analysis can be calculated based on the exported SNP data (see 4.1).

4 Follow-up analyis

4.1 Cluster analysis on SNP data

1. Selecting *File* > *Export to comparison...* (🖫) in the *SNP filtering* window exports the SNP matrix to a new comparison.

In this comparison, the SNP matrix is available as a *character aspect* of the **SNP outbreak** sequence experiment type (see Figure 10).

_	Edit			Statistics Finger		
Exp	_	nents	€□ →	°° + L+ l		
		periment types>		ט	_	🔍 Dend
1		Name		Aspect	•	i
	₹	wgs		<default></default>	^	
	Π.	wgMLST		<all characters=""></all>		
	AC 6 T	denovo		<default></default>		
	82	quality		<all characters=""></all>		
	¥	wgs_TrimmedStats		<default></default>		
	Π.	wgMLST_CallTypes		<all characters=""></all>		
\odot	AC 61	SNP outbreak		SNPs		
	₹	wgsLong		<default></default>		
	AC 6 T	My wgSNP		<default></default>		
					~	The second
<					>	

Figure 10: SNPs character aspect in the *Comparison* window.

We can now create a cluster analysis based on the SNP data, in the same way that a similaritybased clustering is performed in BIONUMERICS:

2. Make sure the **SNP outbreak** experiment is selected in the *Experiments* panel and select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)...*.

Only multi-state similarity coefficients are suitable for clustering of SNP data. The *Categorical (differences)* coefficient treats each different value as a different state, and results in a distance matrix. With the *Scaling factor* one can deal with the hard-coded maximum of 200 that can be calculated for a distance value.

- 3. Select the *Categorical (differences)* coefficient, enter a *Scaling factor* of "100" and press <*Next*> (see Figure 11).
- 4. Select the *Complete linkage* clustering method and press <*Finish*> to calculate the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 12).

5. To view the number of SNPs on the branches, select *Clustering* > *Dendrogram display settings...* (11), and tick the option *Show node information*. Press <*OK*>.

To trace back the number of SNPs from the branches or distance matrix, the displayed values needs to be multiplied with the *Scaling factor* used (here: 100).

6. To clear the selection, press the F4-key.

Comparison settings Page 1 Similarity coefficient		?	×
Character data Character data Character data Pearson correlation Cosine coefficient Canberra metric Bray-Curtis metric Rank correlation Binary Jaccard Dice Simple matching Multi-state Categorical (differences) Categorical (SNPs)	Calculate as distance Scaling factor: 100 Fuzzy logic Tolerance: 0		
Show all.			
	< Back Next >	Can	cel

Figure 11: The categorical similarity coefficient.

The dendrogram contains one well-defined cluster and a number of unrelated strains.

7. Hold the **CTRL**-key and click on this cluster of related strains to select the 10 entries in the database.

This cluster contains all strains with MLST sequence type 22 (see info field "ST info") and corresponds to the strains that were identified as part of the outbreak in the published study (see Figure 12).

Experiment types>	<u>ں</u>	Condrogram	Experiment data	Information fields	Similarities
Name	Aspect	• III 바이 에 에 나 이 나 이 나 이 아이 아	Character name>	₿, ↓↓ \$↑↓₩₩ ↑↓	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
≹ wgs	<default></default>	^	SNP outbreak		
wgMLST	«All Characters»				SNP outbreak
denovo	<default></default>	SNP outbreak (SNPs) 0.487	35514 35514 35514 35287 35287 35287 35642 35644 35679 35644 35679 356515 355515 35555515 3555515 3555515 3555515 355555555		100 80 60 40 20 0
quality	<all characters=""></all>	SNP outbreak (SNPs) 0.487 30 20 10 0		Key ST info	
≩ wgs_TrimmedStats	<default></default>	19.2	G A T G A C A T G G G A C G G A A	ERR103397 1	
wgMLST_CallTypes	<all characters=""></all>	j	GTTGACACTGGACGGT	ERR103395 5	
SNP outbreak	SNPs	0.5	G A A T A C A C T G A A C A G A /	ERR103396 22	
≹ wgsLong	<default></default>	21.2	G A T T A C A C T G A A C A G A /	ERR103400 22	
	>	* 0.0	GATTACACTGAATAGA	ERR103394 22	
		0.0	GATTACACTGAATAGA	ERR103401 22	
ses Local composite datasets		0.1 37.4 0.1	G A T T A C A C T G A A T A G A J	ERR103405 22	
\otimes		0.1	GATTACACTGAATAAA	ERR101900 22	
Name	•	0.4	GATTACACTGAATAGA	ERR101899 22 ERR103403 22	
SNP outbreak (SNPs)		^	GATTACACTGAACAGA	ERR103403 22 ERR103402 22	
		×	AATGGTCCTTAGCAGA	ERR159680 36	
Job log					
⊗ 1 1 ₽					
ize Name					

Figure 12: The Comparison window: wgSNP dendrogram.

8. Save the comparison with the dendrogram by selecting *File* > *Save as...*. Specify a name (e.g. **Neonatal study**) and press <*OK*>.

4.2 Comparison with wgMLST results

For comparison purposes, we will create a cluster analysis for the same entries, but now based on wgMLST data:

- 9. Click on the icon left of **wgMLST** and select the **wgMLST** loci aspect from the drop-down list to visualize the wgMLST allele numbers in the *Experiment data* panel.
- 10. Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)...*, highlight the *Categorical (values)* coefficient and press <*Next*>.
- 11. Leave the default settings enabled and press < *Finish*> to calculate the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 13).

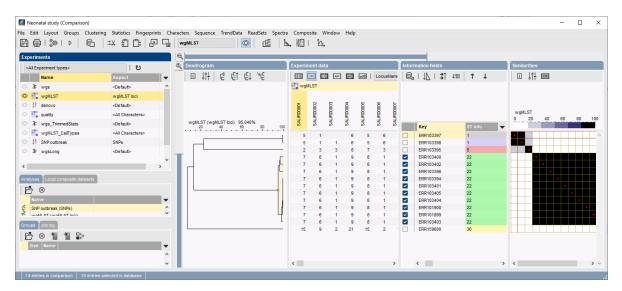


Figure 13: The Comparison window: wgMLST dendrogram.

Via the *Analyses* panel, you can switch back and forth between the two dendrograms. From this visual inspection, it is clear that both analyses correlate very well.

12. Select *Clustering* > *Congruence of experiments...*.

Both experiments show a very high congruence of 95%.

13. Close the *Experiment congruence* window.

4.3 Zooming in on the outbreak

As an alternative to the procedure described previously, a SNP filtering can also be started from the *Comparison* window. We will illustrate this workflow by "zooming in" on the isolates that belong to the outbreak:

14. First, make sure no entries are selected in the *Comparison* window by pressing F4.

- 15. Select the 10 entries that belong to the largest cluster in the dendrogram, e.g. using **Ctrl+click** on the corresponding branch in the dendrogram.
- 16. Return to the *Main* window to create a new comparison for the selected entries (highlight the *Comparisons* panel and select *Edit* > *Create new object...* (+)) or directly using the Alt+C keyboard shortcut.
- A new *Comparison* window pops up with the ten entries that are associated with the outbreak.
 - 17. Select *File* > *Save* (B, Ctrl+S), enter e.g. **Outbreak** as name and press < *OK* >.
 - 18. Click on the sequences in the *Experiment data* panel.
 - 19. Select *Sequence* > *Open SNP window...*
 - 20. Highlight a SNP template from the *SNP analysis* wizard, e.g. *Strict filtering* (the same as we used previously).

This action shows the SNP filtering window again, as discussed previously.

21. Export the SNP matrix back to the comparison via File > Export to comparison... (\mathbb{M}).

This time, a *SNP data name* will be prompted for. Saving SNP matrices under different names will create multiple character aspects for the same sequence type. These aspects are available via the 'Aspect' drop-down list in the *Experiments* panel and can each be used to create cluster analyses from.

- 22. Enter e.g. **SNPs strict** and press <*OK*> in the dialog.
- 23. Return to the **Outbreak** comparison, where the SNP matrix is now displayed.

We can again calculate a dendrogram:

- 24. Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)....*
- 25. Select the *Categorical (differences)* coefficient, specify a *Scaling factor* of "1" and press <*Next*>.
- 26. Check *Complete linkage* for *Method* and press <*Finish*> to calculate the dendrogram.

The resulting dendrogram indicates a sub-structure in the isolates that are associated with the outbreak: two well-defined groups are found (see Figure 14).

4.4 Exporting SNP data

If needed, SNP data can be exported as a character set. We will illustrate this for the **SNPs strict** aspect of **SNP outbreak**:

- 27. In the *Comparison* window, select **SNPs strict** from the 'Aspect' drop-down list next to **SNP outbreak**.
- 28. Select *File* > *Export* > *Export character data...*.
- 29. In the *Export character data* dialog box, make sure *Export mapped values* is checked and press <*OK*>.

The exported SNP matrix will open automatically in MS Excel.

Outbreak (Comparison)						- 0
· · · · ·			naracters Sequence TrendData ReadSets Sp	· · · · · · · · · · · · · · · · · · ·		
] @ ;∞ ⊳ €	∄ ≭X £1 C≩ 6	2 5	SNP outbreak	⊾, ∰] Ŀ,		
xperiments		0				
<all experiment="" types=""></all>	ن	٩		Experiment data	Information fields	Similarities
Name	Aspect	•	3* (3) 13 일 일 (4) 1	Character name>	le, ↓ \$↑ ↓₩₩ ↑ ↓	i ț† III
≩ wgs	<default></default>	^		SNP outbreak		
wgMLST	<all characters=""></all>					SNP outbreak
denovo	<default></default>		SNP outbreak (SNPs strict)	303 303 50566 515151 515151 51565 51515 51565 5515 15555 15755 5555 155555 155555 155555 155555 155555 155555 155555 155555 155555 155555 155555 155555 155555 155555 155555 1555555		100 80 60 40 20 0
quality	<all characters=""></all>		150 100 50 0		Key Organism name ▼ ✓ ERR103400 Staphylococcus aureus	
≵ wgs_TrimmedStats	<default></default>			A A A A C C T T T A T T A T C A	ERR103400 Staphylococcus aureus ERR103402 Staphylococcus aureus	
wgMLST_CallTypes	<all characters=""></all>			AGAACCTTTATTGTCA	ERR103396 Staphylococcus aureus	
SNP outbreak	SNPs strict		ſ	AGCGTAACCGCGGAAG	ERR103394 Staphylococcus aureus	
≹ wgsLong	<default></default>		l _[ERR103404 Staphylococcus aureus ERR103405 Staphylococcus aureus	
	>	× 🗖	-	T G C G T A A C C G C G G A A G	ERR103401 Staphylococcus aureus	
				AGCGTAACCGCGGAAG	ERR101900 Staphylococcus aureus	
lyses Local composite datase	ets				ERR101899 Staphylococcus aureus ERR103403 Staphylococcus aureus	
ጛ⊗						
Name		•				
SNP outbreak (SNPs strict)		^				
		- 11				
		~				
lob log						
ウ⊗ 1 1 ₽⇒						
Size Name		-				
		^				
		~		< >	< >	

Figure 14: Zooming in on the outbreak: two groups.

Alternatively, the data displayed in the *SNP matrix* panel of the *SNP filtering* window can be exported using the column properties button (-) and selecting e.g. *Save content to file*.

Other applications might require the list of SNPs per entry formatted as (pseudo-)sequence:

30. In the *Comparison* window, with the **SNP Strict** aspect still selected, use *File* > *Export* > *Export sequences nonredundant (fasta)*.

The export.txt file that opens is a multi-FASTA file with each row of the SNP matrix represented by a sequence.